

Using Variable Rate Models to Identify Genes Under Selection in Sequence Pairs: Their Validity and Limitations for EST Sequences

Sheri A. Church,¹ Kevin Livingstone,^{1,*} Zhao Lai,¹ Alexander Kozik,² Steven J. Knapp,^{3,†} Richard W. Michelmore,² Loren H. Rieseberg^{1,4}

¹ Department of Biology, Indiana University, Bloomington, IN 47405, USA

² Genome Center and Department of Plant Sciences, University of California, Davis, CA 95616, USA

³ Department of Crop and Soil Science, Oregon State University, Corvallis, OR 97331, USA

⁴ Botany Department, University of British Columbia, Vancouver, B.C. V6T 1Z4, Canada

Received: 10 December 2005 / Accepted: 3 October 2006 [Reviewing Editor: Dr. Rasmus Nielsen]

Abstract. Using likelihood-based variable selection models, we determined if positive selection was acting on 523 EST sequence pairs from two lineages of sunflower and lettuce. Variable rate models are generally not used for comparisons of sequence pairs due to the limited information and the inaccuracy of estimates of specific substitution rates. However, previous studies have shown that the likelihood ratio test (LRT) is reliable for detecting positive selection, even with low numbers of sequences. These analyses identified 56 genes that show a signature of selection, of which 75% were not identified by simpler models that average selection across codons. Subsequent mapping studies in sunflower show four of five of the positively selected genes identified by these methods mapped to domestication QTLs. We discuss the validity and limitations of using variable rate models for comparisons of sequence pairs, as well as the limitations of using ESTs for identification of positively selected genes.

Key words: Selection — EST — Nonsynonymous substitutions — Synonymous substitutions

Introduction

Most molecular studies of positive selection have focused on estimating the type of selection on a gene or gene family of known function (e.g., Wright et al. 2002; O'Connell and McInerney 2005; Strain and Muse 2005) or, more broadly, on genes associated with a particular tissue, such as male reproductive organs (Wyckoff et al. 2000; Swanson et al. 2001, 2003; Torgerson et al. 2002). However large-scale sequencing of genomes, specific genes, and expressed sequence tags (ESTs) now make it feasible to compare substitution rates between species for large numbers of genes from throughout the genome. Currently, only a small number of these whole-genome approaches have been undertaken (e.g., Endo et al. 1996; Swanson et al. 2001; Tiffin and Hahn 2002; Clark et al. 2003). Results from both of these approaches have identified only a small percentage of genes with a molecular signature of positive selection; the remainder exhibit patterns of molecular evolution most consistent with purifying selection or neutral evolution.

One limitation of these whole-genome comparisons is that most only involve comparisons of two sequences (Bishop et al. 2000). Typically, substitution rates are estimated with pairwise maximum likelihood (Goldman and Yang 1994) or approximate (e.g., Nei and Gojobori 1986) methods that average the substitution rates across sites, but averaging may not be appropriate for most genes (e.g., Hughes and Nei 1988; Yang et al. 2000; Bielawski

*Current address: Department of Biology, Trinity University, San Antonio, TX 78212, USA

†Current address: Center for Applied Genetic Technologies, The University of Georgia, Athens, GA 30602, USA

Correspondence to: Sheri A. Church, Department of Biology, George Washington University 340 Lisner Hall, 2023 G Street, N.W., Washington, DC 20052, USA; email: schurch@gwu.edu

and Yang 2001) and has little power (Anisimova et al. 2001). More powerful likelihood methods have been developed that allow substitution rates and selection pressures to vary among sites (variable rate models [Nielsen and Yang 1998; Yang et al. 2000, 2005; Wong et al. 2004]), making it feasible to identify positively selected genes even when most sites are not under selection. These methods have not previously been used for comparisons of sequence pairs because accurate estimates of the strength of selection and the identity of selected sites are sensitive to the number of sequences available for comparison (Anisimova et al. 2001, 2002). However, estimates of the likelihood of a given model of sequence evolution are more robust and less sensitive to sequence number (Anisimova et al. 2001). Thus, comparisons among likelihood estimates of various models provide a reliable means of identifying selection even for pairs of sequences. In this paper, we use both variable rate models that allow selection pressures to vary across sites and rate-averaging models to examine 523 sets of orthologous EST pairs from representatives of the two subfamilies of the Asteraceae: lettuce (*Lactuca sativa* L. and *L. serriola* L.) and sunflower (*Helianthus annuus* L.). Genes under selection within the sunflower or lettuce lineages are identified by comparing the likelihood values of various models, including neutral and selection models, for each gene. To verify the results of the selection analyses, we placed positively selected ESTs from sunflower on a QTL framework map to determine whether they map coincident with QTLs associated with domestication. When possible, sequences were also added from two wild species of sunflower and the selection analyses were repeated. The reliability and applicability of the site-specific selection models for pairs of sequences are discussed in relation to the results from sunflower. Furthermore, we discuss the relative utility of an EST sequencing approach for evolutionary analyses such as those reported here.

Materials and Methods

The EST sequence data were generated as part of the Compositae Genome Project (<http://compngenomics.ucdavis.edu>) and were deposited in our database, which contains ~112,000 individual ESTs sequenced from both sunflower and lettuce (<http://cgpdb.ucdavis.edu>). For sunflower, ~44,000 individual ESTs were sequenced from 10 tissue types (callus, roots, disk and ray flowers, flowers prefertilization, developing kernels, root/shoot chemical induction, roots—environmental stress, shoots—environmental stress, germinating seeds, flowers—environmental stress, hulls) of two *Helianthus annuus* cultivars. The first cultivar, RHA801, is an unbranched confectionary line (Roath et al. 1981), whereas the second, RHA280, is a branched oilseed producing line (Fick et al. 1974). Differences between the cultivars include seed size, seed storage oil production, branching patterns, and disease resistance

(S. Knapp, personal observation). These cultivars have been used in a number of other genetic studies, including a high-density genetic map (Tang et al. 2002).

Individual EST sequences from the two sunflower cultivars were previously assembled into 5504 unique orthologous contiguous sequences (contigs) and 6597 singletons, each of which represents a unique gene (unigene [A. Kozik et al. unpublished]). Of the unigenes with more than one component EST sequence, 2799 are composed of sequences from both sunflower cultivars. A putative reading frame was determined for 2038 of these sunflower unigenes by alignment with the *Arabidopsis thaliana* genome sequence (Arabidopsis Genome Initiative 2000). All sequences, alignments, unigenes (contigs), and results of BLAST searches against *Arabidopsis* are available from the Compositae Genome Project database (<http://cgpdb.ucdavis.edu/>).

The remaining ~68,000 sequences in the dataset are from lettuce. Ten different tissue types (callus, roots, flowers prefertilization, flowers postfertilization, root/shoot chemical induction, roots—environmental stress, shoots—environmental stress, germinating seeds, flowers—environmental stress, dark-grown leaves) from two lettuce species, *Lactuca serriola* 92G489 (a wild lettuce accession) and *L. sativa* cv. Salinas (USDA PI 536851; cultivated crisphead lettuce), were used to generate the ESTs. *Lactuca serriola* may be the wild progenitor of *L. sativa*, and the two are often considered conspecific (Kesseli et al. 1991; for review see Koopman et al. 1998). These species differ in several important characters including bolting, root morphology, leaf shape, and disease resistance (De Vries and Van Raamsdonk 1994). Based on the results of a CAP3 assembly, there are 13,956 singleton unigenes and 8179 unigenes with more than one EST sequence. Sequences from both lettuce species are present in 5341 of these unigenes. The 4641 lettuce unigenes for which putative reading frames have been determined comprise the initial dataset used in the current analyses.

Analyses were performed separately for the sunflower and lettuce genes. A Perl script (available from the authors by request) was written to automate analyses for each sequence set. For sunflower and lettuce unigenes with more than one EST sequence from a taxon, a majority consensus sequence was constructed for the lineage using modified versions of the bioperl (Stajich et al. 2002) modules “consensus_jupac” and “consensus_string” (available from the authors by request). The majority consensus rule was applied only where ESTs had sequence information for a position (missing data were ignored), allowing single EST sequences to extend the majority consensus sequence. At positions where EST reads overlapped but there was not a majority consensus across sequences, International Union of Pure and Applied Chemistry ambiguity codes were used. The translated region of each consensus sequence was then extracted using “extractseq” from the European Molecular Biology Open Software Suite (Rice et al. 2000).

From the sunflower and lettuce datasets, sequence pairs were excluded from analysis based on poor alignment scores, stop codons within the putative reading frames, or very short consensus sequences for either or both of the lineages (9% removed from sunflower, 5% removed from lettuce). Unigenes with very short (<100 codons) or very similar (distance <0.11) sequences were excluded from analysis based on the simulation parameters of Anisimova et al. (2001).

For the remaining sequences, the nucleotide composition of the translated regions was analyzed using the Phylogenetic Analysis Using Maximum Likelihood (PAML) package (version 3.14; Yang 1997). The CODEML module in PAML uses maximum likelihood methods to estimate the likelihood scores of various models for a given gene. Using a maximum likelihood approach can account for transition/transversion rate biases, codon usage bias, and multiple substitutions (Yang and Nielsen 1998).

The likelihood score of each of six analyses was computed for each pair of consensus sequences using CODEML. Four analyses implemented site-specific models (M0, M1a, M2a, M3 [Nielsen and

Yang 1998; Yang et al. 2000]) and two analyses implemented the pairwise comparison method (runmode -2). Three models (M0 and two pairwise models, runmode -2) were implemented that average synonymous and nonsynonymous substitution rates across sites. This results in a single selection pressure across codons (d_N/d_S). In the first pairwise model (runmode -2), d_N/d_S was estimated from the data. The second pairwise model (runmode -2) assumed d_N/d_S to be fixed at 1, a neutral model. In model M0, d_N/d_S is estimated from the data. Three site models allowing selection pressure to vary across codons (M1a, M2a, M3) were also implemented (Nielsen and Yang 1998; Yang et al. 2000, 2005; Wong et al. 2004). These models allow the nonsynonymous substitution rate to vary over sites, whereas the synonymous substitution rate remains constant. For these models, the average d_N/d_S across the gene is a function of codon specific substitution rates (ω_n) and the proportion of codons with each substitution rate (p_n). In the nearly neutral model, M1a, two site classes are allowed. The first class estimates the nonsynonymous rate, but it is constrained to be between 0 and 1 (ω_0 at proportion p_0 of sites). The second class of sites constrains the nonsynonymous substitution rate to be equal to the synonymous substitution rate ($\omega_1 = 1$ at proportion p_1 of sites; $d_N/d_S = \omega_0 * p_0 + \omega_1 * p_1$). The selection model, M2a, incorporates a third class of sites (ω_2) that estimates an unconstrained nonsynonymous substitution rate. The final model, M3, is a discrete model with three classes ($K = 3$) in our analyses. This model allows nonsynonymous substitution rates to vary across codons, estimating three categories of substitution rates (ω_0 , ω_1 , and ω_2 , in proportions p_0 , p_1 and p_2 , respectively; $d_N/d_S = \omega_0 * p_0 + \omega_1 * p_1 + \omega_2 * p_2$).

Several of these models are nested with one another, allowing their likelihood values to be directly compared using a likelihood ratio test (LRT). In this test, the likelihood value of a simpler (null) model ($-\ln L_1$) is compared to the likelihood value of a more general alternative model ($-\ln L_2$) by taking twice their difference and comparing this value with a chi-square distribution. Specifically, the likelihood values of the two pairwise analyses (d_N/d_S estimated vs. constrained; runmode -2) can be compared and have 1 degree of freedom. The selection model (M2a) can be compared to the neutral model (M1a) with 2 degrees of freedom (Yang et al. 2000) and the discrete model (M3; three nonsynonymous substitution rate classes) can be compared to model M0, which contains only one class of nonsynonymous substitution rates (df = 3). For the pairwise models (runmode -2), if the constrained model ($d_N/d_S = 1$) is rejected in favor of the unconstrained model that estimates d_N/d_S and $d_N/d_S > 1$, then this is a test for positive selection. Similarly, if the selection model (M2a) is a significantly better fit to the data than the neutral model (M1a) and contains a class of sites with $\omega > 1$, this also constitutes a significant test for positive selection (Yang et al. 2000). A significant LRT between the M3 and the M0 models is support for variable substitution rates among codons.

To test the reliability of the LRT between variable selection pressure models, we asked whether the ESTs that appeared to be under positive selection in sunflower were correlated with domestication traits. This was accomplished by PCR-amplifying the positively selected ESTs using the "touch-down" cycling program described by Burke et al. (2002). Loci that successfully amplified were genotyped on a WAVE dHPLC (denaturing high-performance liquid chromatograph) following the methods of Lexer et al. (2003). MAPMAKER 3.0/EXP (Lincoln et al. 1992) was used to place the ESTs on an existing QTL framework map for wild X domesticated sunflower (Burke et al. 2002; Lai et al. 2005). We also searched much smaller EST databases for two wild sunflower species, *H. argophyllus* (sister to *H. annuus*) and *H. paradoxus* (hybrid derivative of *H. annuus* and *H. petiolaris*), for orthologous sequences to the positively selected sunflower genes. Sequences were found for only one positively selected gene, and the analyses were repeated to determine if the results were consistent with those obtained from sequence pairs.

Results

Fifty-six genes were identified as having experienced positive selection in either the sunflower or the lettuce lineages (Tables 1 and 2). Although the functions of most of the identified genes under positive selection are unknown, several exhibit homology to known genes. Inferred functions of these genes in sunflower include the regulation of transcription, a dehydration induced protein, and an oxidation enzyme. In lettuce, the inferred functions include the regulation of cell division and gene expression, RNA unwinding and splicing, calcium and protein transport, anthocyanin biosynthesis, floral development, and construction of the cytoskeleton (see database). The genes under positive selection were found to occur in every tissue type included in the EST libraries except from pre-fertilized *Helianthus* flowers, for which very few ESTs were sequenced. Moreover, most of the selected genes were expressed in multiple tissues. There was thus no statistical difference among tissue categories in the occurrence of selected genes. Unpublished mapping studies in sunflower also show that there is no apparent clustering of these genes within the genome.

Sunflower

Of the original 2038 unigenes with a predicted reading frame and at least one sequence from both the oilseed and the confectionary sunflower taxa, only 1479 had more than 100 codons. Fifteen percent (224) of these genes had sufficient numbers of substitutions ($0.11 < \text{distance} < 2$) for further analyses. For these 224 genes, the average number of codons is 215 (range, 100–694) and the average distance between sequences is 0.363. The current analyses identified 11 genes (4.9%) under positive selection based on significant LRTs and at least one class of substitutions with $\omega > 1$ (Table 1). In 10 of these cases, both the selection model (M2a) and the discrete model (M3) fit the data significantly better than either the neutral model (M1a) or the rate-averaging model (M0). The LRT between the pairwise models (runmode -2) was significant for five genes; however, the estimated d_N/d_S value in the pairwise comparison was not > 1 . The remaining 213 sunflower genes appear to be evolving under neutral conditions or purifying selection.

Amplification products were generated for all 11 positively selected genes in sunflower, however, dHPLC patterns were too complex for some of them to allow confident mapping. Five genes were successfully mapped and are positioned on the wild X domesticated QTL map of Burke and Rieseberg (2002) as follows: gene 502 (between marker 328 and marker 1043 on linkage 8), gene 566 (between marker

Table 1. Genes determined to be under positive selection in sunflower cultivars

Contig ID	No. codons	Distance	Models	d_N/d_S	LRT	Significance
260	218	0.335	M0:M3	0.8236	7.743	
			M1a:M2a		6.983	*
			Runmode -2	0.8245	0.072	
502	240	0.343	M0:M3	0.3726	15.204	**
			M1a:M2a		12.814	**
			Runmode -2	0.3666	2.737	
566	239	0.122	M0:M3	0.333	11.825	**
			M1a:M2a		7.946	*
			Runmode -2	0.3261	4.419	*
774	180	0.224	M0:M3	0.3114	11.871	**
			M1a:M2a		7.326	*
			Runmode -2	0.3144	2.970	
1353	260	0.114	M0:M3	0.1084	10.125	*
			M1a:M2a		6.466	*
			Runmode -2	0.0631	12.057	***
1548	180	1.120	M0:M3	0.615	12.110	**
			M1a:M2a		10.513	**
			Runmode -2	0.6509	0.531	
1827	251	0.397	M0:M3	0.3818	13.661	**
			M1a:M2a		11.379	*
			Runmode -2	0.3838	1.868	
2337	228	0.197	M0:M3	0.5289	9.554	*
			M1a:M2a		7.801	*
			Runmode -2	0.5213	0.926	
2604	223	0.402	M0:M3	0.3174	11.721	**
			M1a:M2a		7.066	*
			Runmode -2	0.3014	5.997	*
2815	426	1.274	M0:M3	0.1755	34.316	***
			M1a:M2a		21.338	***
			Runmode -2	0.2334	5.006	*
2816	267	0.983	M0:M3	0.1836	36.255	***
			M1a:M2a		28.080	***
			Runmode -2	0.183	9.080	**

Note. All contigs from sunflower are preceded by “QH_CA_” in the Compositae database. Gene function is hypothesized based on *Arabidopsis* genome annotation. The d_N/d_S ratio estimates are based on models M0 and the runmode -2 pairwise comparison only.

* Significant at $\alpha = 0.05$; ** significant at $\alpha = 0.01$; *** significant at $\alpha = 0.001$.

258 and marker 343 on linkage 16), gene 1548 (between marker 995 and marker 388 on linkage 13); gene 2337 (between marker 727 and marker 561 on linkage 17), and gene 2816 (above marker 388 on linkage 13). Four of these positions are very close to domestication QTLs. Specifically, gene 566 (putative transcription factor, based on BLAST results) maps coincident with a QTL controlling the number of heads per branch; gene 1548 (putative protein) maps to the same position as a QTL affecting leaf shape; gene 2337 (putative protein) maps coincident with a QTL affecting days to flower, plant height, number of leaves, and peduncle length; and gene 2816 (putative b-zip transcription factor) maps to a QTL controlling disk diameter, ray number, and achene width. Sequence from gene 2816 was also found in the database for *Helianthus paradoxus* and used in an analysis combining the three genes with variable substitution rate models. The results of this analysis are consistent with the results presented here, with a single class of amino acids having an elevated substitution rate

($\omega = 5.22$) and the LRT between the selection (M2a) and neutral (M1a) models being significant.

Lettuce

After removing genes with too few codons, 3755 lettuce genes were retained. Only 299 (8%) of these genes showed sufficient levels of divergence ($0.11 < \text{distance} < 2$) for further analyses. On average, these genes had estimated genetic distances of 0.36 and 259 codons (range, 101–733). Of these, 45 (17.4%) show evidence of positive selection (Table 2), while the remaining 254 appear to be either evolving neutrally or under selective constraint.

For all positively selected genes except one (see below), the selection model (M2a) fit the data significantly better than the neutral model (M1a) (Table 2). For 39 of the genes under positive selection, the discrete model (M3) was also significantly better than the rate-averaging model (M0). Only nine genes

Table 2. Genes determined to be under positive selection in lettuce species

Contig ID	No. codons	Distance	Models	d_N/d_S	LRT	Significance
95	151	0.798	M0:M3	0.305	11.777	*
			M1a:M2a		8.309	*
			Runmode -2	0.300	1.303	
227	212	0.360	M0:M3	0.145	15.772	**
			M1a:M2a		10.894	**
			Runmode -2	0.145	3.897	*
1183	204	0.213	M0:M3	0.441	21.015	***
			M1a:M2a		15.422	***
			Runmode -2	0.303	3.920	*
1233	222	0.338	M0:M3	1.103	46.735	***
			M1a:M2a		41.363	***
			Runmode -2	1.105	0.013	
1816	209	0.130	M0:M3	0.367	20.756	***
			M1a:M2a		17.574	***
			Runmode -2	0.361	0.944	
1822	228	0.341	M0:M3	0.778	11.276	*
			M1a:M2a		10.096	**
			Runmode -2	0.781	0.078	
2120	199	0.183	M0:M3	0.855	36.823	***
			M1a:M2a		31.866	***
			Runmode -2	0.849	0.035	
2138	418	1.931	M0:M3	0.638	11.727	*
			M1a:M2a		9.195	*
			Runmode -2	0.518	0.249	
2148	212	0.497	M0:M3	1.150	22.969	***
			M1a:M2a		19.418	***
			Runmode -2	1.154	0.030	
2318	471	0.271	M0:M3	0.054	19.155	***
			M1a:M2a		17.000	***
			Runmode -2	0.099	3.460	
2410	542	1.220	M0:M3	0.170	17.665	**
			M1a:M2a		9.767	**
			Runmode -2	0.171	10.284	**
2681	194	0.477	M0:M3	0.327	15.009	**
			M1a:M2a		12.119	*
			Runmode -2	0.322	1.922	
2822	547	0.283	M0:M3	1.613	9.358	
			M1a:M2a		8.768	*
			Runmode -2	1.482	0.148	
3239	188	1.502	M0:M3	0.323	16.448	**
			M1a:M2a		13.702	**
			Runmode -2	0.251	2.623	
4186	149	0.211	M0:M3	0.341	19.247	***
			M1a:M2a		15.624	***
			Runmode -2	0.580	0.145	
4352	274	0.245	M0:M3	1.007	13.994	**
			M1a:M2a		12.754	**
			Runmode -2	0.908	0.007	
4604	164	1.188	M0:M3	0.516	27.456	***
			M1a:M2a		23.559	***
			Runmode -2	0.524	1.131	
4750	279	0.218	M0:M3	0.570	27.950	***
			M1a:M2a		22.473	***
			Runmode -2	0.565	0.351	
4852	300	0.628	M0:M3	0.918	44.955	***
			M1a:M2a		37.969	***
			Runmode -2	0.919	0.015	
4886	137	0.159	M0:M3	0.513	21.814	***
			M1a:M2a		18.910	***
			Runmode -2	0.488	0.301	
5005	203	1.107	M0:M3	1.082	17.611	**
			M1a:M2a		13.766	**
			Runmode -2	1.046	0.013	

(Continued)

Table 2. Continued

Contig ID	No. codons	Distance	Models	d_N/d_S	LRT	Significance
5028	413	0.208	M0:M3	0.467	13.570	**
			M1a:M2a		11.218	**
			Runmode -2	0.868	0.014	
5544	247	0.144	M0:M3	0.916	10.856	*
			M1a:M2a		8.792	*
			Runmode -2	0.934	0.007	
5618	157	0.499	M0:M3	2.037	17.682	**
			M1a:M2a		17.27	**
			Runmode -2	1.972	0.451	
5649	323	0.16304	M0:M3	5.284	0.000	
			M1a:M2a		3.906	
			Runmode -2	5.342	3.945	*
5721	376	0.430	M0:M3	0.068	19.877	***
			M1a:M2a		13.419	**
			Runmode -2	0.077	9.540	**
5806	515	0.303	M0:M3	0.170	16.237	**
			M1a:M2a		11.056	**
			Runmode -2	0.171	7.864	**
6144	294	0.425	M0:M3	0.581	15.824	**
			M1a:M2a		15.331	***
			Runmode -2	0.555	0.892	
6195	215	0.392	M0:M3	0.972	14.086	**
			M1a:M2a		13.583	**
			Runmode -2	0.946	0.008	
6212	477	0.886	M0:M3	0.305	10.356	*
			M1a:M2a		7.633	*
			Runmode -2	0.303	2.730	
6213	202	0.649	M0:M3	0.396	11.625	*
			M1a:M2a		11.228	**
			Runmode -2	0.246	1.442	
6371	248	0.435	M0:M3	0.442	12.872	*
			M1a:M2a		9.163	*
			Runmode -2	0.438	0.839	
6480	144	0.261	M0:M3	0.261	11.593	*
			M1a:M2a		7.808	*
			Runmode -2	0.275	2.625	
6567	310	0.677	M0:M3	0.111	12.624	*
			M1a:M2a		9.534	**
			Runmode -2	0.112	7.630	**
6568	213	0.518	M0:M3	0.112	19.266	***
			M1a:M2a		14.046	***
			Runmode -2	0.112	7.564	**
6627	232	0.504	M0:M3	0.511	8.867	*
			M1a:M2a		8.357	*
			Runmode -2	0.466	0.416	
6637	569	0.208	M0:M3	0.270	11.930	*
			M1a:M2a		10.591	**
			Runmode -2	0.271	5.038	*
6686	191	0.331	M0:M3	0.197	7.168	*
			M1a:M2a		7.161	**
			Runmode -2	0.197	2.960	
6804	446	0.278	M0:M3	1.220	17.421	**
			M1a:M2a		15.959	**
			Runmode -2	1.224	0.098	
6867	198	0.403	M0:M3	0.589	12.073	*
			M1a:M2a		10.301	**
			Runmode -2	0.575	0.357	
6919	408	0.172	M0:M3	3.660	4.568	*
			M1a:M2a		6.688	*
			Runmode -2	3.763	2.188	
7650	223	0.310	M0:M3	0.341	9.455	*
			M1a:M2a		6.959	*
			Runmode -2	0.340	3.011	

(Continued)

Table 2. Continued

Contig ID	No. codons	Distance	Models	d_N/d_S	LRT	Significance
7892	154	0.654	M0:M3	0.566	12.259	*
			M1a:M2a		7.249	*
			Runmode -2	0.611	0.840	
7968	154	0.900	M0:M3	0.769	19.561	***
			M1a:M2a		15.208	***
			Runmode -2	0.769	0.245	
8053	208	0.447	M0:M3	0.402	19.488	***
			M1a:M2a		16.171	***
			Runmode -2	0.403	2.017	

Note. All lettuce contigs are preceded by “QG_CA_” in the Composite database. Differences in log-likelihood values represent the comparison between model M2 and model M1 (M2 rows), M3 and M0 (M3 rows), and ML and ML-neutral models (ML rows).

* Significant at $\alpha = 0.05$; ** significant at $\alpha = 0.01$; *** Significant at $\alpha = 0.001$.

had a significant LRT between pairwise models (runmode -2), with one of these genes having an estimated $d_N/d_S > 1$. In only one case did the pairwise analyses identify a gene (lettuce 5649) under positive selection that was not identified by the more complex variable rate models. It is interesting to note that for this same gene, the variable rate model (M3) had the same likelihood as the rate-averaging model (M0), which had a d_N/d_S value of 5.284, suggesting that in this case the variable rate models were not necessary to identify positive selection.

Discussion

Evolutionary analyses of shotgun EST sequence data from closely related sunflower and lettuce taxa have identified 56 genes whose coding regions appear to be under positive selection. The majority of genes were identified by implementing site-specific variable substitution rate models. In sunflower, these results were strengthened by correlations with domestication QTLs and by sequence comparisons with closely related species. These results suggest that more complex models can successfully identify positively selected genes from pairs of sequences. Although we successfully identified a number of genes under positive selection, only a small fraction of the sequenced genes showed high enough levels of diversity to be used in such analyses. This indicates that although evolutionary analyses of ESTs are informative, a large number of sequences must be obtained to ensure a sufficient sample size for subsequent evolutionary comparisons.

Alternate Models of Selection

In our analyses, 55 genes under positive selection were identified based on a significant LRT between the site-specific variable substitution rate selection

model (M2a) and the site-specific neutral model (M1a). Only a subset of these genes, 15 (27%), had a significant LRT between the pairwise models (runmode -2). Reliance on the latter model alone would have resulted in underestimates of the proportion of genes under positive selection, overlooking evolutionarily important genes.

However, it is important to consider the limits of the variable rate models as well, particularly when pairs of sequences are compared. Simulation studies have shown that increasing the number of sequences in a comparison bolsters the accuracy of the variable rate models, particularly in calculations of the strength of selection (ω), the proportion of sites under selection (p), and subsequent identification of specific sites under selection (Anisimova et al. 2001, 2002). It is likely that estimates of these parameters in the current study are not accurate and, therefore, were not reported. Fortunately, estimates of the likelihood of a given substitution rate model for a gene are accurate and the LRT is reliable, even for smaller datasets (Anisimova et al. 2001). Briefly, simulations indicate that the LRT is a conservative test even when sequences are short or divergence is low (Anisimova et al. 2001), such as in our study. With few sequences, the LRT loses power, but accuracy is not affected (Anisimova et al. 2001), making it more likely that our results misidentified positively selected genes as being neutral rather than the reverse. Even given these limitations, the LRT was able to identify a number of candidate positively selected genes. Furthermore, in sunflower, results from subsequent analyses and mapping studies were consistent with the results from the LRT. Our results thus demonstrate the value of using models that incorporate variable substitution rates across codons, even when only two sequences are available for comparison. The assumption of positive selection on individual genes should be verified using mapping or sequencing

techniques such as those used with the sunflower dataset. However, the methods presented here make these other techniques much more feasible due to the much smaller pool of target genes.

Genes Under Selection

It is important to note that while our study examines agricultural lineages that have been subject to artificial selection, our estimates of the number of positively selected genes (5–17%) are similar to those estimates obtained from noncultivated species (0%–11% [e.g., Swanson et al. 2001; Tiffin and Hahn 2002; Clark et al. 2003]). Due to the short time since the domestication of these species, the allelic differences within positively selected genes may predate domestication and have been either sorted among domesticated lines (sunflower) or sampled from the wild progenitor (lettuce). Sampling of other wild and cultivar populations will help us to better pinpoint the location and timing of the positive selection.

Mapping Studies in Sunflower

Although four of the positively selected genes in sunflower are correlated with domestication QTLs, much additional work will be required to test whether they are causally related to these phenotypic differences. Full-length cDNA sequences are required to more precisely infer what the function of these genes might be. Two of them are unknown proteins. The other two have inferred functions (b-zip transcription factor and DNA binding protein) that are consistent with the phenotypic changes correlated with them, but the functional categories are too broad to be informative. In addition to the sequencing, fine-mapping is needed to verify the correlation and transgenic complementation and/or RNAi mediated by virus-induced silencing (Baulcombe 1999; Chuang and Meyerowitz 2000) will be required to demonstrate function.

Evolutionary Analyses of ESTs

The tremendous number of ESTs currently being generated for a variety of different plant and animal species present the tantalizing potential to identify genes contributing to species differences. However, coding sequences often are not sufficiently divergent to test for positive selection in closely related taxa. In our analyses, for example, only 10% of the unigenes with more than 100 codons were sufficiently divergent for analyses of substitution rates and selection. The proportion of genes with sufficient divergence for analysis is likely to increase for comparisons of more distantly related taxa, but the fraction of analyzable

genes may still remain small due to other limiting factors, such as lack of reliable BLAST hits and hence reliable translated regions and reading frames. Furthermore, evolutionary analyses of EST data rely on the identification of orthologous sequences from unique genes. In our analyses, we carefully examined unigene sets for any indication of paralogy such as multiple patterns of substitution at a given gene within taxa. However, for many genes, only a single EST sequence was available from each lineage. In this case, paralogues could be misidentified as orthologues and any signature of selection could be due to divergence among genes rather than among lineages. Thus, it is important to verify the orthology of these sequences with further molecular analyses such as direct sequencing.

Assuming that orthologous gene copies are correctly identified, another concern is that most EST sequences have untranslated regions that are not informative for tests of positive selection. In our analyses, the untranslated region averaged 100 bp. As a result, many unigenes were too short for rigorous analysis of positive selection. Given that most ESTs (or the unigenes derived from them) do not cover the entire gene, some genes under positive selection are likely to be missed. Moreover, most of the unigenes in both lettuce and sunflower are represented by single EST sequences and may contain uncorrected sequencing errors. These factors will reduce the number of sequences that can be tested for positive selection and will bias downward estimates of the proportion of positively selected genes. Cumulatively, these results indicate that although evolutionary analyses of ESTs are productive, a large number of sequenced genes are needed to ensure a sufficient sample size for evolutionary analyses.

Conclusions

To our knowledge, this is the first attempt to use variable substitution rate models to compare sequence pairs. The current analyses are also among the first to test for positive selection across a large number of genes isolated from throughout the genome as well as across several tissue types. The results have identified 56 genes that are under positive selection in cultivated taxa of sunflower and lettuce. This corresponds to ~11% of the analyzed genes, which is in accordance with previous studies of positive selection. In sunflower, several of these positively selected genes map coincident with QTL involved in domestication. While we were able to identify a significant number of positively selected genes, we have also identified several limitations to the use of EST sequences for similar evolutionary analyses. In particular, the short lengths of many

unigenes and low divergence levels between taxa excluded many genes from these analyses. As a result, less than 0.5% of the genes in the original unigene set could be confidently identified as experiencing positive selection.

Acknowledgments. We thank Kent Bradford and Rick Kesseli for assisting in development of the EST database as well as the members of the Compositae Genome Group for valuable discussions concerning this project. We also thank J. P. Bielawski for helpful discussions concerning the implementation of maximum likelihood models for pairwise data. The manuscript was greatly improved by comments from two anonymous reviewers and the associate editor, Rasmus Nielsen. This work was supported by grants from the U.S. Department of Agriculture (00-32100-9609 to R.W.M., S.J.K., and L.H.R. and NRI 2001-35301-09971 to K.L.) and a National Science Foundation Postdoctoral Research Fellowship in Biological Informatics to S.A.C. (DBI-0204160).

References

- Anisimova M, Bielawski JP, Yang Z (2001) Accuracy and power of likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol* 18:1585–1592
- Anisimova M, Bielawski JP, Yang Z (2002) Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Mol Biol Evol* 19:950–958
- Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815
- Baulcombe DC (1999) Fast forward genetics based on virus-induced gene silencing. *Curr Opin Plant Biol* 2:109–113
- Bielawski JP, Yang Z (2001) Positive and negative selection in the *DAZ* gene family. *Mol Biol Evol* 18:523–529
- Bishop JG, Dean AM, Mitchell-Olds T (2000) Rapid evolution in plant chitinases: Molecular targets of selection in plant-pathogen coevolution. *Proc Natl Acad Sci USA* 97(10):5322–5327
- Burke JM, Tang S, Knapp SJ, Rieseberg LH (2002) Genetic analysis of sunflower domestication. *Genetics* 161:1257–1267
- Chuang CF, Meyerowitz EM (2000) Specific and heritable genetic interference by double-stranded RNA in *Arabidopsis thaliana*. *Proc Natl Acad Sci USA* 97:4985–4990
- Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, Ferreira S, Wang G, Zheng X, White TJ, Sninsky JJ, Adams MD, Cargill M (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* 302:1960–1963
- De Vries IM, Van Raamsdonk LWD (1994) Numerical morphological analysis of lettuce cultivars and species (*Lactuca* sect. *Lactuca*, Asteraceae). *Pl Syst Evol* 193:125–141
- Endo T, Ikeo K, Gojobori T (1996) Large-scale search for genes on which positive selection may operate. *Mol Biol Evol* 13:685–690
- Fick JD, Zimmer DE, Kinman ML (1974) Registration of six sunflower parental lines. *Crop Sci* 14:912
- Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* 11:725–736
- Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex loci reveals overdominant selection. *Nature* 335:167–170
- Kesseli RV, Ochoa O, Micheltore RW (1991) Variation at RFLP loci in *Lactuca* spp. and the origin of cultivated lettuce. *Genome* 54:430–436
- Koopman WJM, Guetta E, van de Wiel CCM, Vosman B, van den Berg RG (1998) Phylogenetic relationships among *Lactuca* (Asteraceae) species and related genera based on ITS-1 DNA sequences. *Am J Bot* 85:1517–1530
- Kozik A, Micheltore RW, Knapp SJ, et al. (2002) Lettuce and sunflower ESTs from the Compositae Genome Project. Available at: <http://www.cgdb.ucdavis.edu/>
- Lai Z, Livingstone K, Zou Y, Church SA, Knapp SJ, Andrews J, Rieseberg LH (2005) Identification and mapping of SNPs from ESTs in sunflower. *Theor Appl Genet* 111:1532–1544
- Lexer C, Lai Z, Rieseberg LH (2004) Candidate gene polymorphisms associated with salt tolerance in wild sunflower hybrids; implications for the origin of *Helianthus paradoxus*, a diploid hybrid species. *New Phytol* 161:225–233
- Lincoln S, Daly M, Lander E (1992) Constructing genetic maps with MAPMAKER/EXP 3.0. Technical Report. Whitehead Institute, Cambridge, MA
- Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 3:418–426
- Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148:929–936
- O'Connell MJ, McInerney JO (2005) Gamma chain receptor interleukins: evidence for positive selection driving the evolution of cell-to-cell communicators in the mammalian immune system. *J Mol Evol* 61:608–619
- Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16:276–277
- Roath WW, Miller JF, Gulya TJ (1981) Registration of RHA 801 sunflower germplasm. *Crop Sci* 21:479
- Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigan C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka ED, Wilkinson M, Birney E (2002) The Bioperl Toolkit: Perl modules for the life sciences. *Genome Res* 12:1161–1168
- Strain E, Muse SV (2005) Positively selected sites in the *Arabidopsis* receptor-like kinase gene family. *J Mol Evol* 61:325–332
- Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF (2001) Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. *Prog Natl Acad Sci* 98:7375–7379
- Swanson WJ, Nielsen R, Yang Z (2003) Pervasive adaptive evolution in mammalian fertilization proteins. *Mol Biol Evol* 20:18–20
- Tang S, Yu JK, Slabaugh MB, Shintani DK, Knapp SJ (2002) Simple sequence repeat map of the sunflower genome. *Theor Appl Genet* 105:1124–1136
- Tiffin P, Hahn MW (2002) Coding sequence divergence between two closely related plant species: *Arabidopsis thaliana* and *Brassica rapa* spp. *pekinensis*. *J Mol Evol* 54:746–753
- Torgerson DG, Kulathinal RJ, Singh RS (2002) Mammalian sperm proteins are rapidly evolving: evidence for positive selection in functionally diverse genes. *J Mol Biol Evol* 19:1973–1980
- Wong WSW, Yang Z, Goldman N, Nielsen R (2004) Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* 168:1041–1051
- Wright SI, Lauga B, Charlesworth D (2002) Rates and patterns, of molecular evolution in inbred and outbred *Arabidopsis*. *Mol Biol Evol* 19:1407–1420
- Wyckoff GJ, Wang W, Wu C-I (2000) Rapid evolution of male reproductive genes in the descent of man. *Nature* 403:304–309
- Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS* 13:555–556

- Yang Z, Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J Mol Evol* 46:409–418
- Yang Z, Nielsen R, Goldman N, Pedersen AMK (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449
- Yang Z, Wong WSW, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol* 22:1107–1118